Royal Meteorological Society

Validation of the forecast skill of the Global Modeling and Assimilation Office Observing System Simulation Experiment

N. C. Privé,[a]* R. M. Errico[a] and K.-S. Tai[b]

[a]*Goddard Earth Sciences Technology and Research Center, Morgan State University, Baltimore, MD, USA*
[b]*Science Systems and Applications Inc., Greenbelt, MD, USA*

*Correspondence to: N. C. Privé, Global Modeling and Assimilation Office, NASA/GSFC Code 610.1, Greenbelt, MD 20771, USA. E-mail: nikki.prive@nasa.gov

A global Observing System Simulation Experiment (OSSE) framework has been developed at the National Aeronautics and Space Administration Global Modeling and Assimilation Office (NASA/GMAO). The OSSE uses a 13-month integration of the European Centre for Medium-Range Weather Forecasts (ECMWF) operational model as the Nature Run, and the Goddard Earth Observing System version-5 (GEOS-5) forecast model with Gridpoint Statistical Interpolation (GSI) data assimilation as the forecast model. Synthetic observations have been developed with correlated observation errors to replicate the observing network from 2005–2006.

The performance of the GMAO OSSE in terms of forecast skill and observation impacts is evaluated against real observational data for the same period. Metrics of anomaly correlation of 500 hPa geopotential and root-mean-square error of temperature and wind fields for 120 h forecasts are calculated for once-daily forecasts from July 2005, and an adjoint is used to measure observation impacts of different data types. The forecast skill of the OSSE is found to be slightly higher than for real data, with smaller observation impacts overall, possibly due to insufficient model error in the OSSE. While there is similar relative ranking of observation impact for most data types in the OSSE compared with real data, for individual satellite channels the agreement is not as good. Some caveats and difficulties of using the OSSE system are discussed along with recommendations of how to avoid potential pitfalls when performing OSSEs. Copyright © 2012 Royal Meteorological Society

Key Words: data assimilation; observation impact

Received 20 March 2012; Revised 6 July 2012; Accepted 31 July 2012; Published online in Wiley Online Library

Citation: Privé NC, Errico RM, Tai K-S. 2012. Validation of the forecast skill of the Global Modeling and Assimilation Office Observing System Simulation Experiment. *Q. J. R. Meteorol. Soc.* DOI:10.1002/qj.2029

1. Introduction

Observing System Simulation Experiments (OSSEs) are modelling studies used to evaluate the potential improvement in numerical weather forecasts due to the introduction of a new observing system. An OSSE can be performed before the new observing system is developed, and many different variations and variables can be tested to determine the optimal design and deployment of the new system. As demonstrated by Errico *et al.* (2007), OSSEs are also powerful tools for investigating the behaviour of data assimilation systems (DASs).

An OSSE consists of three components: an extended free run of a forecast model that represents ‘truth’ for the OSSE, referred to as the Nature Run (NR); a full set of synthetic observations for all observation types currently ingested into operational DAS with values derived from the NR fields; and a second forecast model and DAS that are used with the synthetic observations to generate the experimental forecasts. The model used to generate the NR and that used for the second forecast model should be different to avoid the identical-twin problem of insufficient model error. The NR should have higher resolution than the second forecast model, and should accurately represent the atmospheric features that are to be studied in the experimental forecasts. Synthetic observations should have temporal and spatial distribution similar to real-world observations, and should have appropriate added errors.

An OSSE has been developed at the National Aeronautics and Space Administration (NASA) Global Modeling and Assimilation Office (GMAO). This OSSE uses a 13-month free run of a 2006 version of the European Centre for Medium-Range Weather Forecasts (ECMWF) operational numerical weather prediction model as the NR. The Goddard Earth Observing System Model, Version-5 (GEOS-5) (Rienecker *et al.*, 2008) forecast model with Gridpoint Statistical Interpolation (GSI) data assimilation (Kleist *et al.*, 2009) is used for the forecast experiment model.

It is important to verify that the OSSE behaves in a way that is sufficiently similar to the real world for the results to be of value. This verification process has several stages: evaluation of the NR behaviour in comparison to the real atmosphere, calibration of the synthetic observations, and testing of the OSSE system as a whole. The behaviour of the entire GMAO OSSE system is addressed in this article; evaluation of the ECMWF NR was addressed by Reale *et al.* (2007), Masutani *et al.* (2007) and McCarty *et al.* (2012), and calibration of the synthetic observations was discussed by Errico *et al.* (2012).

Previous OSSEs have often used simple calibration methods, or were uncalibrated. One method of calibration that has been used is the data-denial experiment (Masutani *et al.*, 2006), where observation impact is measured for both synthetic and real data. Observation minus analysis and observation minus background statistics have also been used for calibration of OSSEs (Stoffelen *et al.*, 2006). However, calibration has generally been measured using few metrics, simply to verify adequate performance of the OSSE system, and not as a method for improving the OSSE behaviour. The approach taken for calibration of the GMAO OSSE differs in that the synthetic observations were iteratively adjusted through repeated calibration tests to improve the performance of the OSSE system.

This article describes the performance of the GMAO OSSE in terms of forecast metrics and observation impacts. The development of the OSSE is detailed in section 2, and analysis of the experiments is described in section 3. Section 4 gives a discussion of the results.

2. Set-up

A detailed description of the components of the GMAO OSSE is given by Errico *et al.* (2012); a brief overview is given here.

The NR was generated by ECMWF as part of a larger joint OSSE cooperative effort using the operational forecast model version c31r1. The run was performed from 01 May 2005 to 31 May 2006 at T511 resolution with 91 vertical levels. The only forcings were sea surface temperature and sea ice; these were taken from archival data from the same time period.

Synthetic observations were generated based on the temporal and spatial distribution of real archived observations during the period of the NR. Conventional data types were created by interpolating the NR fields to the time and location of recorded observations. Radiance observations were created using the Community Radiative Transfer Model (CRTM; Han *et al.*, 2006) to generate brightness temperatures calculated from the NR fields, including a simplified treatment of cloud effects using the NR high-, mid-, and low-cloud fractions. The archived radiance data were partially thinned prior to creation of the synthetic observations to reduce the computational cost.

Errors were added to the synthetic data to simulate a combination of observation error and representativeness error. Some representativeness error is also intrinsic to the synthetic observations due to the difference in resolution between the NR and the forecast model. Uncorrelated errors were added to all observation types and an additional component of correlated errors was added to some types. Vertically correlated errors were added to conventional sounding data types, horizontally correlated errors were added to AMSU, HIRS, and MSU* observations, channel correlated errors were added to AIRS†, and both vertically and horizontally correlated errors were added to satellite wind observations. No correlation of errors was applied between different data types, and observation bias was not added. The observation errors were calibrated to match the error correlation, analysis increment, and observation minus forecast statistics seen for real data; Errico *et al.* (2012) give details.

The GEOS-5 atmospheric model version 5.7.1, and GSI data assimilation system were selected as the forecast model for experiments. The GEOS-5 model is described in detail by Rienecker *et al.* (2008) and the GSI is described by Kleist *et al.* (2009). The model resolution used was 0.5° in latitude and 0.625° in longitude with 72 vertical layers.

In order to evaluate the performance of the OSSE system, the GEOS-5 model was cycled for a period from 14 June 2005 to 31 July 2005. Only analyses and forecasts from July 2005 were evaluated as June was treated as a spin-up period for the system. For each day of the period, 120 h forecasts were launched from 0000 UTC. This process was performed twice: (i) using archived real observations from this period (Control), and (ii) using the synthetic observations with calibrated errors added (OSSE). The analysis and forecast error statistics were compared to evaluate how well the OSSE emulates the real world. An adjoint of the GEOS-5 model and analysis system was used to estimate the impacts of individual observation types on the 24 h forecasts (Gelaro and Zhu, 2009).

3. OSSE evaluation

The metrics that are used to investigate a new observing system with an OSSE include observation impacts relative to other data types, forecast anomaly correlations, and root-mean-square (RMS) forecast errors verified against the NR. When evaluating the performance of the OSSE, these metrics should be tested in comparison to real-world metrics in order to validate the results of the OSSE. However, the corresponding true state of the real atmosphere is not known, so validation of the OSSE must be performed by calculating these metrics with verification against the analysis

*Advanced Microwave Sounding Unit; High-Resolution Infrared Sounder; Microwave Sounding Unit.

†Atmospheric InfraRed Sounder.

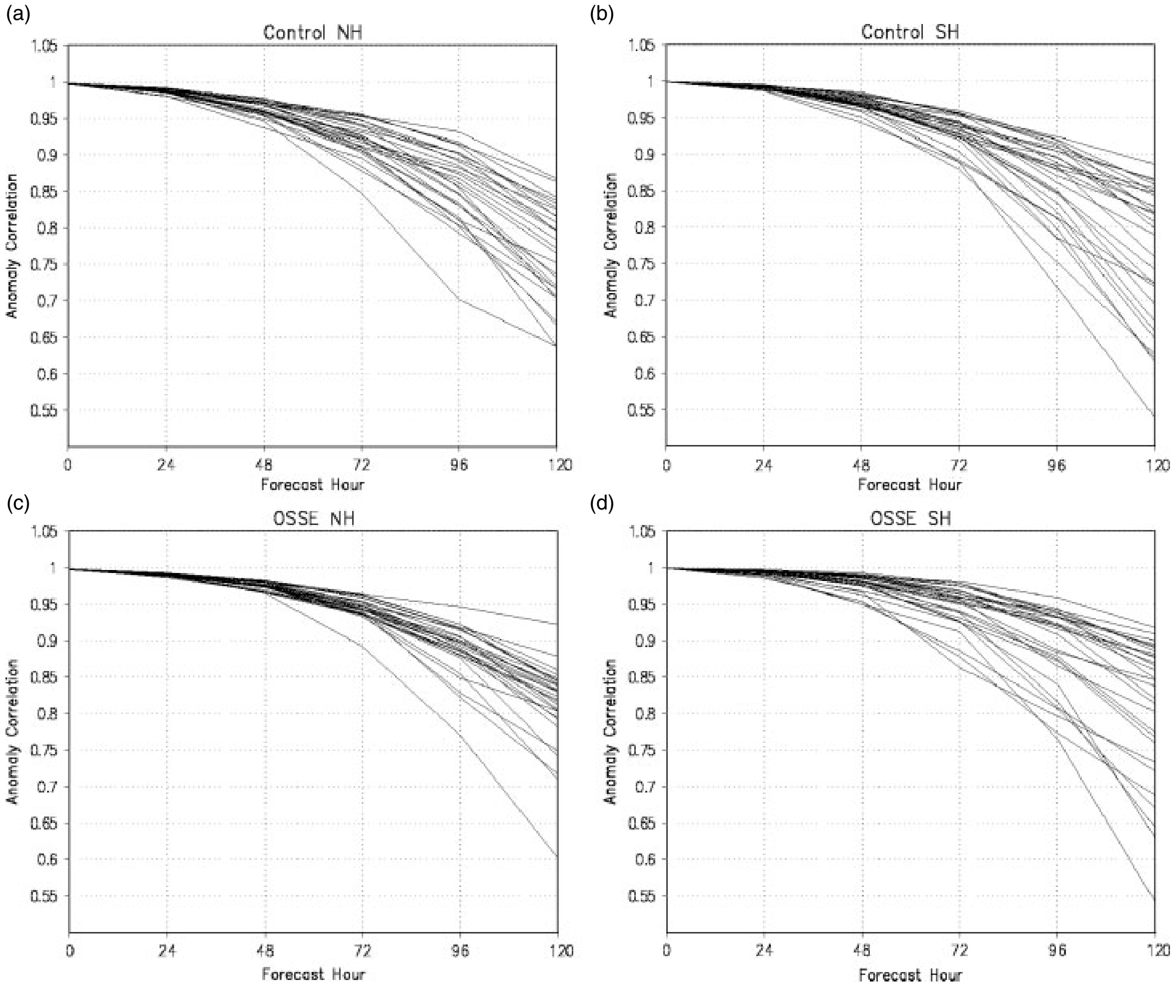

Figure 1. Anomaly correlations for 500 hPa geopotential for forecasts from 2 July to 30 July 2005 for (a, b) Control, and (c, d) OSSE over the (a, c) Northern Hemisphere midlatitudes and (b, d) Southern Hemisphere midlatitudes. The thin black lines represent individual forecasts.

Table 1. Mean anomaly correlation for 24 h and 120 h forecasts over 2 July–30 July 2005 calculated using a Fisher *z* transform over the Northern Hemisphere (NH; 30–90°N), and Southern Hemisphere (SH; 30–90°S).

	24 h		120 h	
	NH	SH	NH	SH
Control	0.988	0.992	0.771	0.777
OSSE	0.991	0.995	0.816	0.826

state rather than against the NR. It is also important that the OSSE not be ‘over-tuned’ and that correct results are not due to incorrect behaviour. This can be determined by examination of analysis increment statistics and other measures of how the data assimilation system handles the observation information.

The GMAO OSSE has undergone evaluation of certain analysis statistics during calibration and validation of the synthetic observations (Errico *et al.*, 2012). Observation innovation (observation minus background) and analysis increment (analysis minus background) metrics in the troposphere were used to tune synthetic observation errors and verify the OSSE behaviour. It was found that the analysis statistics were significantly improved by the addition of correlated observation errors for radiance and satellite wind observations. The fully tuned OSSE was found to have observation innovation and analysis increment statistics that were quite similar to the statistics of assimilation of real data, although the variance of the analysis increment was slightly smaller in the OSSE.

3.1. *Forecast statistics*

OSSEs are often used to evaluate the potential impact of new observing systems on multi-day forecasts; therefore the OSSE forecasts should behave similarly to operational forecasts. A frequent concern is that the forecast model and the NR model will be too similar, resulting in insufficient forecast error in the OSSE.

Anomaly correlations of 500 hPa geopotential are frequently used to measure the accuracy of forecasts in the midlatitudes. The anomaly correlations for the Northern and Southern Hemispheres (NH, SH) are shown as functions of forecast hour in Figure 1 and the means of the 24 h and 120 h anomaly correlations are listed in Table 1. As the anomaly correlation distribution is not normal, a Fisher *z* transform (e.g. Buizza, 1997) is used to calculate the mean. At 120 h the mean of the forecast skill is higher in the OSSE than in the Control in both hemispheres, but there is no significant difference in the forecast skill between OSSE and Control at 24 h, using a Mann–Whitney *U* test (Mann and Whitney, 1947) to determine significance at the 95% level.

The spread of forecast skill in Figure 1 shows significant differences between the Control and OSSE in the NH. The

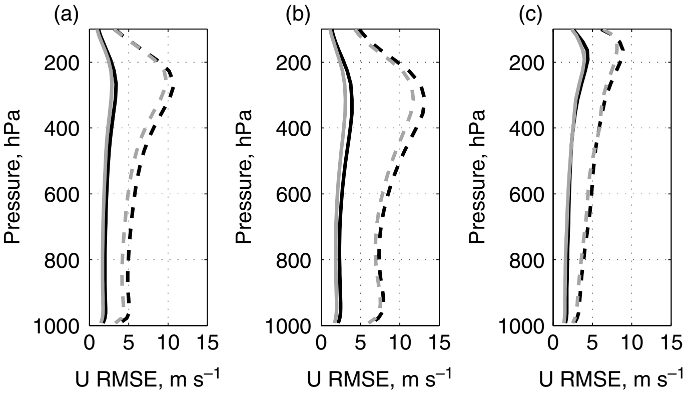

Figure 2. Root mean square 24 h (solid lines) and 120 h (dashed lines) forecast error verified against analysis for the Control (black lines) and OSSE (grey lines), areal mean for zonal wind (m s^{-1}). (a) 30–90°N, (b) 30–90°S, and (c) 30°N–30°S.

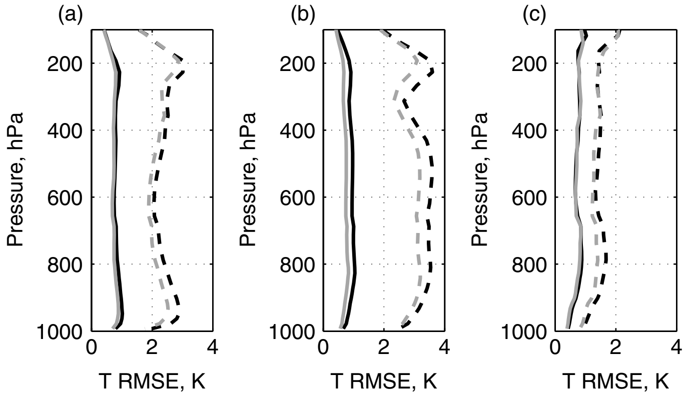

Figure 3. As Figure 2 but for *T* (K).

OSSE has a tight cluster of forecasts near the mean with only a few low-skill forecasts, while the Control does not show clustering and has more lower-skill forecasts. For the SH, there is a better match between the OSSE and Control forecast skill spread, with some clustering at higher skill along with numerous lower-skill forecasts. There can be considerable differences in month to month predictability and forecast skill due to the synoptic state of the atmosphere, so that forecasts over several additional Julys would be needed to determine if differences in the forecast skill spread of the OSSE is systematically different from that of the Control. Unfortunately, only a single year is available for the current NR.

Root-mean-square error (RMSE) between the forecast and verifying analysis is also a useful metric for evaluating forecast skill, particularly in the Tropics. Figures 2 and 3 illustrate the 24 h and 120 h forecast RMSE for zonal wind and temperature as functions of model level. The Tropics show very good agreement between the OSSE and Control cases in terms of RMSE for specific humidity (not shown), wind, and temperature for the 24 h forecasts. By 120 h, the OSSE has smaller RMSE temperature than Control in the lower and middle troposphere Tropics, but shows good agreement for temperature at upper levels and for wind throughout the tropical troposphere.

In the midlatitudes, agreement in RMSE between the Control and OSSE is better in the NH than the SH. At the 24 h forecast, the NH temperature RMSE is 4–12% higher in the Control with wind error 10–15% higher than in the OSSE. The wind error remains 10–15% higher for the Control at 120 h, with the temperature error difference increasing slightly to 8–12%. In the SH, the RMSE mismatch between the Control and OSSE decreases from 20–25% for both wind and temperature at 24 h to 10% for temperature and 4–12% for wind at 120 h.

The forecast error can be separated into monthly mean error (forecast 'bias') and anomaly error (forecast error minus bias). The area-averaged global root mean square mean error (RMSME) and root mean square anomaly error (RMSAE) are calculated and shown in Figure 4 for temperature and Figure 5 for zonal wind. The RMSME at the 24 h forecast for zonal wind is 20% smaller for the OSSE than Control in the SH and 15% smaller in the NH. The discrepancy diminishes slightly in the midlatitudes for the 120 h forecast, with 7–12% smaller RMSME difference in the NH zonal wind and 7–18% smaller RMSME difference in the SH. This suggests that the climatology of the GEOS-5 model zonal wind is closer to the NR climatology than to the real climatology in the midlatitudes. The magnitude of the RMSME is smaller than the magnitude of the RMSAE, so this difference in climatologies plays only a minor role in the total forecast error of the Extratropics. In the Tropics, the zonal wind RMSME increases from 3–11% at the 24 h forecast to 9–12% error by 120 h.

The RMSME for temperature is 8–21% smaller in the SH in the OSSE than the Control at the 24 h forecast, decreasing to 8–11% at 120 h. However, larger temperature RMSME is found in the OSSE in the NH upper troposphere, where the error is almost 40% higher than the Control near 200 hPa at the 120 h forecast. In the Tropics, the upper-tropospheric RMSME is also larger in the OSSE than the Control, although the tropospheric vertical mean difference is only 1% at 120 h.

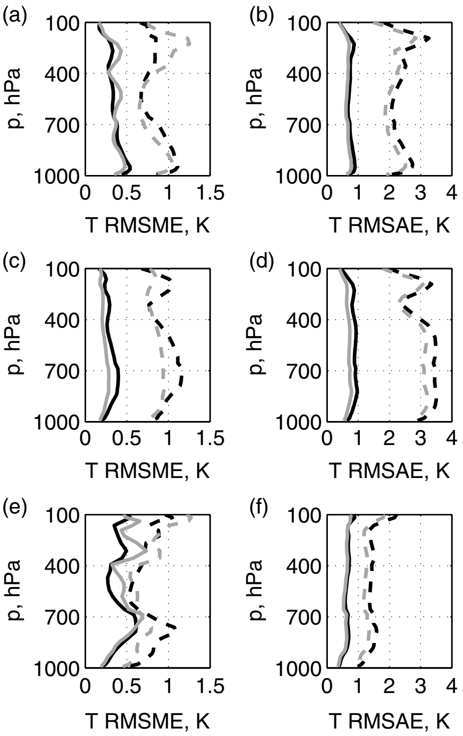

Figure 4. Comparison of 24 h (solid lines) and 120 h forecast error (dashed lines) versus analysis for the Control (black lines) and OSSE (grey lines) for temperature (K). (a, c, e) show RMSME, and (b, d, f) show RMSAE. (a, b) 30–90°N; (c, d) 30–90°S; (e, f) 30°S–30°N.

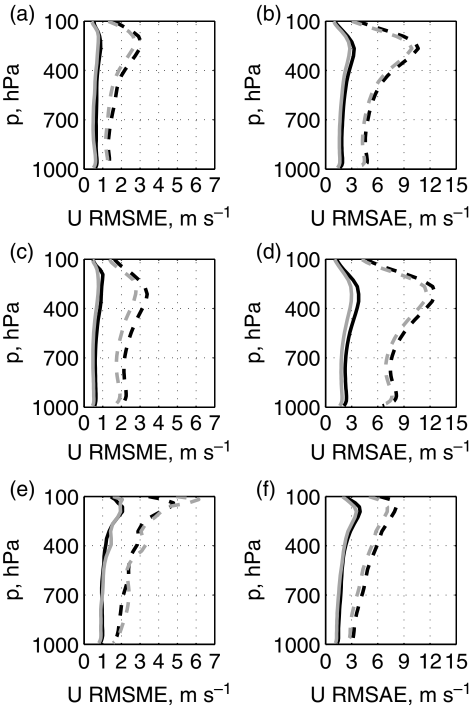

Figure 5. As Figure 4, but for zonal wind u (m s^{-1}).

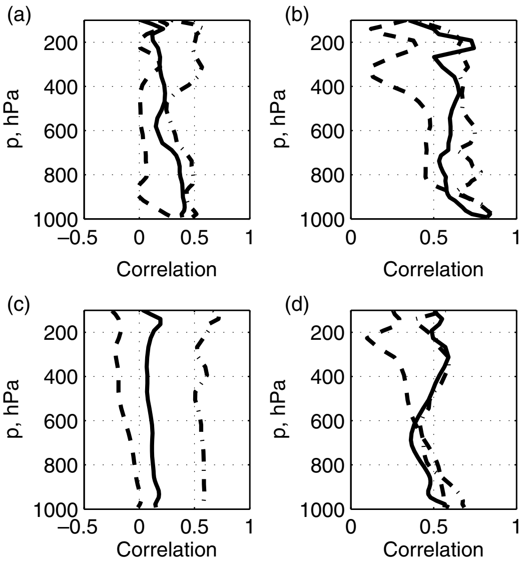

Figure 6. Spatial correlation of the 120 h forecast error between Control and OSSE: solid line, 30–90°N, dashed line 30–90°S, dot-dash line 30°S–30°N. (a, c) monthly mean forecast error; (b, d) root- monthly mean-square anomaly forecast error. (a, b) T (K); (c, d) u (m s^{-1}). Note the different x-axis ranges.

The spatial correlation between the OSSE and Control mean error fields is shown in Figure 6(a, c), and an illustration of the mean error distribution for the Control and OSSE temperature at 356 hPa is shown in Figure 7(a, b). The largest correlations are found in the Tropics, with near zero or anticorrelation seen in the SH midlatitudes. This implies that the difference in the mean error between Control and OSSE seen in Figures 4 and 5 is a matter of the spatial pattern of climatology in the midlatitudes but that differences in mean error in the Tropics may be dominated by disparities in the magnitude of climatological features between the NR and the real world.

The RMSAE magnitude is greater in the Control than in the OSSE for all regions, with largest discrepancy in the SH midlatitudes (Figures 4 and 5); because RMSAE is non-negative, the correlation will be increased. The correlation of the OSSE and Control standard deviation of anomaly error is highest in the Tropics and NH midlatitudes, as seen in Figure 6(b, d), with maps of the RMSAE shown in Figure 7(c, d). The correlation of anomaly error in the SH is in the range 0.1–0.4 in the middle and upper troposphere while the NH correlation range is 0.5–0.7. The higher correlations in the NH may be due to the strong influence of continents on the storm tracks in the NH which is lacking in the SH.

The forecast errors at 24 h and 120 h are decomposed into power spectra using a Fourier transform along three latitudes: 50°N, 50°S and 0°N. The spectra for each latitude are averaged over 25 forecasts from 2 July to 26 July, and are shown in Figures 8 and 9. The best agreement between the Control and OSSE cases is seen in the Tropics, where the spectra are very similar at 24 h and 120 h forecasts. Larger discrepancies are seen in the midlatitudes, particularly at wavenumbers less than 10, where the OSSE has lower spectral amplitude than the Control in the 24 h forecast for both temperature and zonal wind. There is also a greater bias in the Control midlatitudes, particularly for zonal wind (not shown). By the 120 h forecast, the low-wavenumber error spectra amplitude is very similar overall for the Control and OSSE. The RMS error has not begun to saturate in the midlatitudes at 120 h, but error growth in the Tropics has slowed by this time (not shown).

At high wavenumbers, a variety of behaviours are seen in the spectra of forecast error. For temperature at 50°N, the OSSE error has lower amplitude at 24 h but similar amplitude to Control at 120 h, while at 50°S, the amplitude

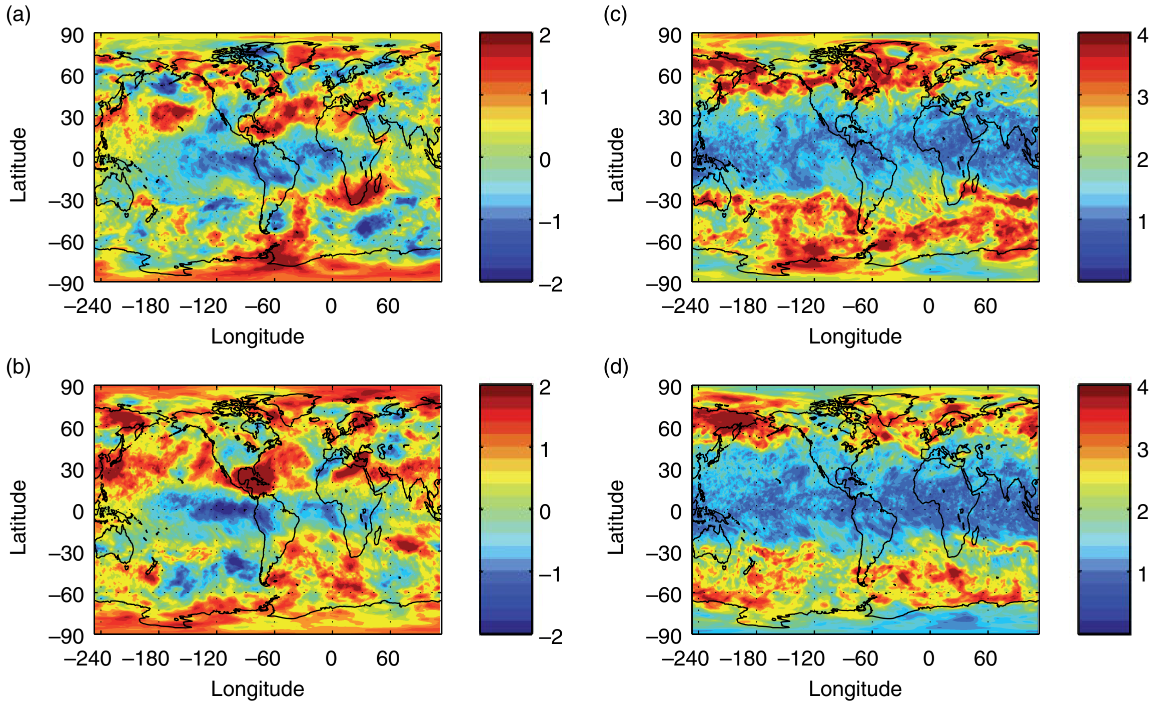

Figure 7. 120 h forecast error compared with verifying analysis field for T (K) at the 356 hPa sigma level. (a, b) monthly mean forecast error; (c, d) root-monthly mean-square anomaly forecast error. (a, c) Control; (b, d) OSSE.

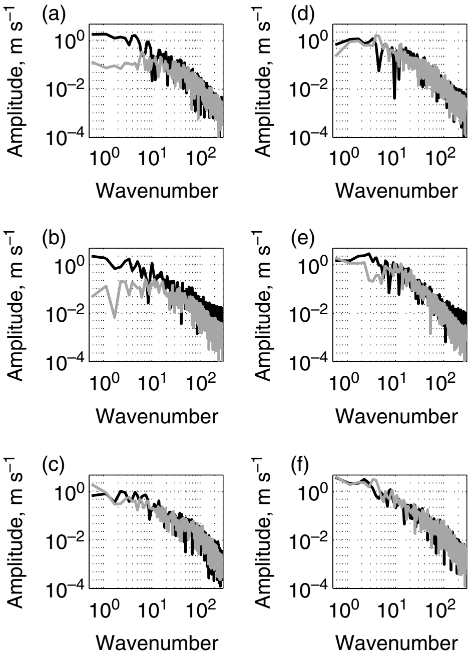

Figure 8. Frequency spectra of spatial forecast error compared to analysis, zonal wind (m s^{-1}) for the Control (black line) and OSSE (grey line): (a, b, c) 24 h forecast, and (d, e, f) 120 h forecast. (a, d) 50°N; (b, e) 50°S; (c, f) 0°N.

of the OSSE is similar to the Control at 24 h but significantly higher in the OSSE at 120 h. For zonal wind at 50°N, the high-wavenumber error amplitude is similar in the OSSE and Control at 24 h and 120 h, but at 50°S the amplitude is larger in the Control at both 24 h and 120 h.

The lower amplitude of low-wavenumber errors seen in the OSSE 24 h forecast compared to the Control forecast could have several sources such as lack of bias added to the synthetic radiance observations and insufficient model error.

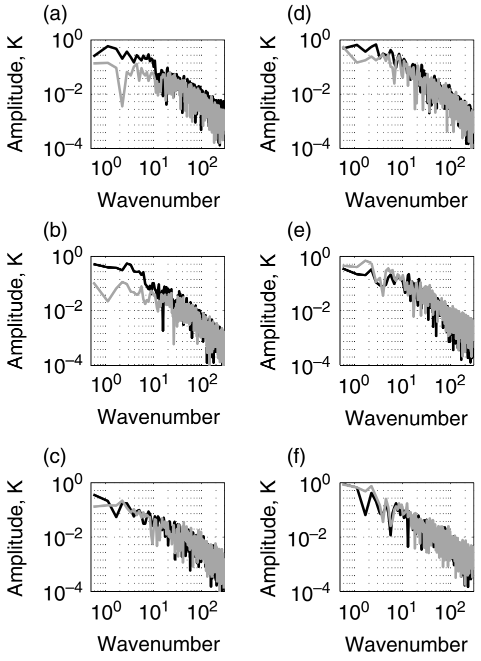

Figure 9. As Figure 8, but for temperature (K).

Only the uncorrected part of bias in the real observations would impact the Control analysis, but if the bias of an observation type changes rapidly in comparison to the 30 day e-folding time-scale of the relaxation of the bias correction routines, significant bias may go uncorrected during data assimilation. This type of bias is not explicitly added to synthetic observations because it is not well understood or quantified in real data.

If the error growth rate of large-scale errors is persistently slower in the OSSE compared to the real world, significant discrepancies in low-wavenumber error spectra would be expected at both 24 h and 120 h forecasts. Since a significant difference between the OSSE and Control is seen in the early forecast but not the extended forecast, this might indicate that error growth is too slow during the initial forecast but that the time-scale for error growth may shorten after the 24 h forecast.

The generally good agreement between the OSSE and Control error spectra at 120 h for wavenumbers 1–50 lends confidence to the performance of the OSSE in terms of forecast skill of medium-range forecasts. Because the discrepancies between the OSSE and Control at high wavenumber are not consistent, it is not obvious how to assign a likely source of the differences. For example, insufficient representativeness error in the OSSE might be expected to reduce the amplitude of errors at high wavenumbers, but this is not consistently noted in the results.

3.2. *Observation impacts*

An adjoint has been developed for GEOS-5 that uses simplified dry physics; the adjoint is fully described and evaluated in Gelaro and Zhu (2009). The full physics trajectories are calculated for 24 h forecasts from both the analysis and corresponding background at a reduced horizontal resolution of 1° with 72 vertical levels, then adjoint forecasts are calculated for these trajectories using a simplified dry physics at the same resolution. The metric selected for the adjoint is the dry energy norm (Talagrand, 1981; Errico, 2000) which is influenced by the temperature, surface pressure, and horizontal wind fields. The forecast error is calculated for the dry energy norm, and the GSI adjoint (Zhu and Gelaro, 2008) is run at the same resolution as the model adjoint. The observation impacts are derived from the GSI adjoint results and the observation innovations.

The adjoint results have been shown (Gelaro and Zhu, 2009) to be comparable to observing system experiment results for data denial tests. However, the adjoint results omit some observation impacts, in particular for instruments that measure humidity –this is due to both the choice of a dry metric and the absence of moist processes in the adjoint itself. The adjoint results are a measure of how much ‘work’ observations contribute toward improving the analysis and short-term forecast. If the quality of the background field is high, there will be only a limited amount of possible improvement, and the observation impact will be small.

In order to compare observation impacts between the Control and OSSE, the forecasts at 24 h were verified against the corresponding analysis fields. Adjoint calculations were performed on 27 forecasts starting at 0000 UTC from 2 July 2005 to 31 July 2005, with three missing forecasts on 3, 17, and 25 July due to data archiving problems.

Observation impacts derived from the adjoint are shown in Figure 10 for conventional and radiance data types, with a negative (positive) impact indicating a reduction (increase) in error of the 24 h forecast due to the observation. The total observation impact per data type is displayed rather than the impact per observation; as the data counts in the OSSE are comparable to the real data Control by design, comparison of the per observation impacts do not contain additional information. RAOB[‡] and AMSU-A observations have the greatest impacts on the 24 h forecast, in agreement with previous studies of operational numerical weather prediction systems (Zapotocny *et al.*, 2008). The very small impact for AMSU-B and conventional humidity data is likely due to the choice of a dry metric for the adjoint.

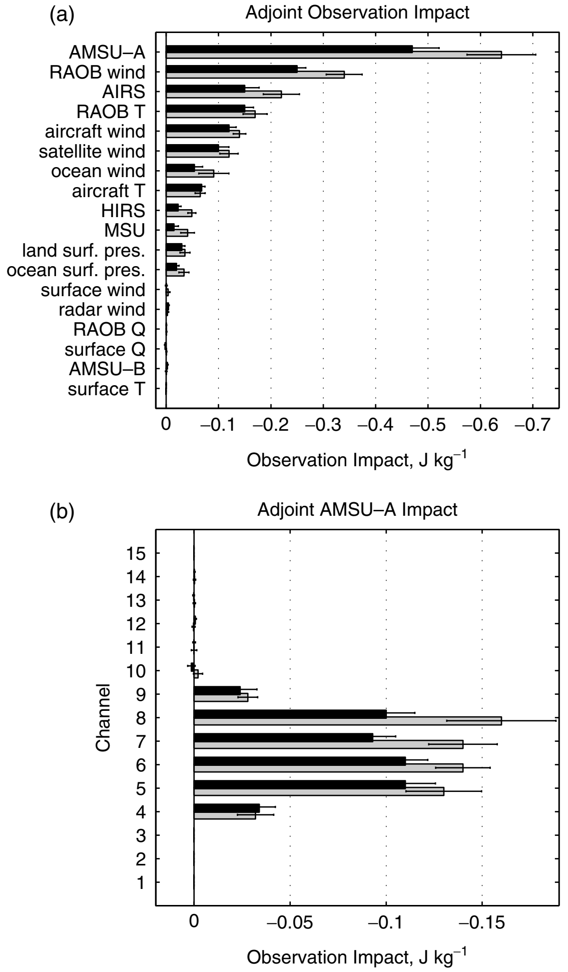

Figure 10. Observation impact from adjoint calculations using dry error energy norm, J kg^{-1} for the Control (grey bars) and OSSE (black bars). Error bars indicate 95% confidence interval. (a) radiance and conventional observation types; (b) AMSU-A impact per channel.

The relative magnitudes of observation impacts between data types is fairly consistent between the Control and OSSE cases, although those for the OSSE are approximately 30% smaller than for the Control case for most data types. Given that the RMSE of the 24 h forecast is lower in the OSSE than in the Control as described in section 3.1, it is likely that insufficient model error in the OSSE results in smaller background error and thus lower observation impacts, as discussed further in section 4.

Adjoint results can also be narrowed to evaluate particular regions, levels, or satellite channels. Figure 10(b) shows the observation impact for AMSU-A calculated per channel. Impact is low in the OSSE for the significant channels with the exception of channel 4. The relative impact of channels 5 to 8 in the OSSE is not consistent with the Control observation impact ranking. As the discrepancies between the NR and the real world climatologies are particularly large near the surface and in the upper atmosphere, it is expected that those satellite channels which feel surface effects or peak in the stratosphere or higher may not be represented well in the OSSE. Channels 5 and 6 for AMSU-A have weighting functions with significant contribution near the

[‡] Rawindsonde Observation.

surface, while channel 7 has a small contribution from the surface and channel 8 is not impacted by the lowest levels of the troposphere; this surface influence on some of the AMSU-A channels may contribute to the discrepancy in relative channel impact.

The channel breakdown of observation impacts for AIRS is shown in Figure 11 for the AIRS channels used by the GEOS-5 GSI with peak weighting functions below 100 hPa. Similar to the AMSU-A results, the relative impacts of the AIRS channels in the OSSE shows gross correspondence to real data, but the channels with strongest impact are considerably weaker in the OSSE, and some channels (for example, channel 1565) even have the wrong sign. There does not seem to be any correspondence between the height of the AIRS weighting function peaks and the agreement between the OSSE and Control case observation impact.

4. Discussion

The ultimate goal of OSSE development is the creation of a synthetic system with behaviour identical to the real world; in practice, this goal is unattainable (Rosenblueth and Wiener, 1945). A more practical objective is to achieve reasonable agreement with the real world for key properties. These key properties include both metrics that are intended for experimental testing in the OSSE system as well as for the intrinsic behaviours that influence these metrics, so that 'the right results are seen for the right reasons' and not due to overtuning of the system. This is not a trivial process, and the selection of appropriate metrics should be approached with care.

Some aspects of the OSSE system can be modified or tuned to adjust the behaviour of the system, while other aspects are difficult or impossible to change. The components of the OSSE that are simplest to change are the synthetic observations and their explicitly-added error characteristics. Improvements to the methods used to generate the synthetic observations, for example increasing the sophistication of the treatment of clouds, may change the analysis increment or observation impact of a data type. However, certain aspects of the OSSE are much more difficult to modify, such as the forecast model error. If the GEOS-5 forecast model behaviour is more akin to the ECMWF model than to the real world, model error will be insufficient and the effects will be seen both in the quality of the background field and in the extended forecast skill.

The results of the experiments performed thus far are not sufficient to determine whether the forecast skill in the OSSE system is within the same range as forecast skill for real observations. It is unclear what the month-to-month variation of forecast skill is in either the Control or OSSE; in order to determine this, a full year or more of forecasts might be necessary at great computational expense. Forecast skill is impacted not only by analysis and model error, but also by the intrinsic predictability of the atmospheric state. It is possible that differences in the predictability of the NR state compared to the real atmosphere during 2005–2006 contribute to the discrepancies in forecast skill. The role of model error can be quantified in an OSSE setting as it is possible to initialize the forecast model with the NR fields for a 'perfect' initial condition. Further exploration of these issues is planned for future investigations.

Given that the forecast skill metrics of the OSSE may differ from those of the real world, what are the limitations of performing experiments? For the GMAO OSSE, forecast skill in the Tropics is commensurate with the real world when measured with broad-brush metrics such as RMSE, but forecast skill in the midlatitudes is slightly higher in the OSSE. Since the forecast skill has been rather insensitive to observation error in the OSSE during the calibration process, it is likely that achieving a better match of midlatitude forecast skill between the OSSE and real world would be laborious and difficult if not impossible with the constraints of the current NR and forecast model.

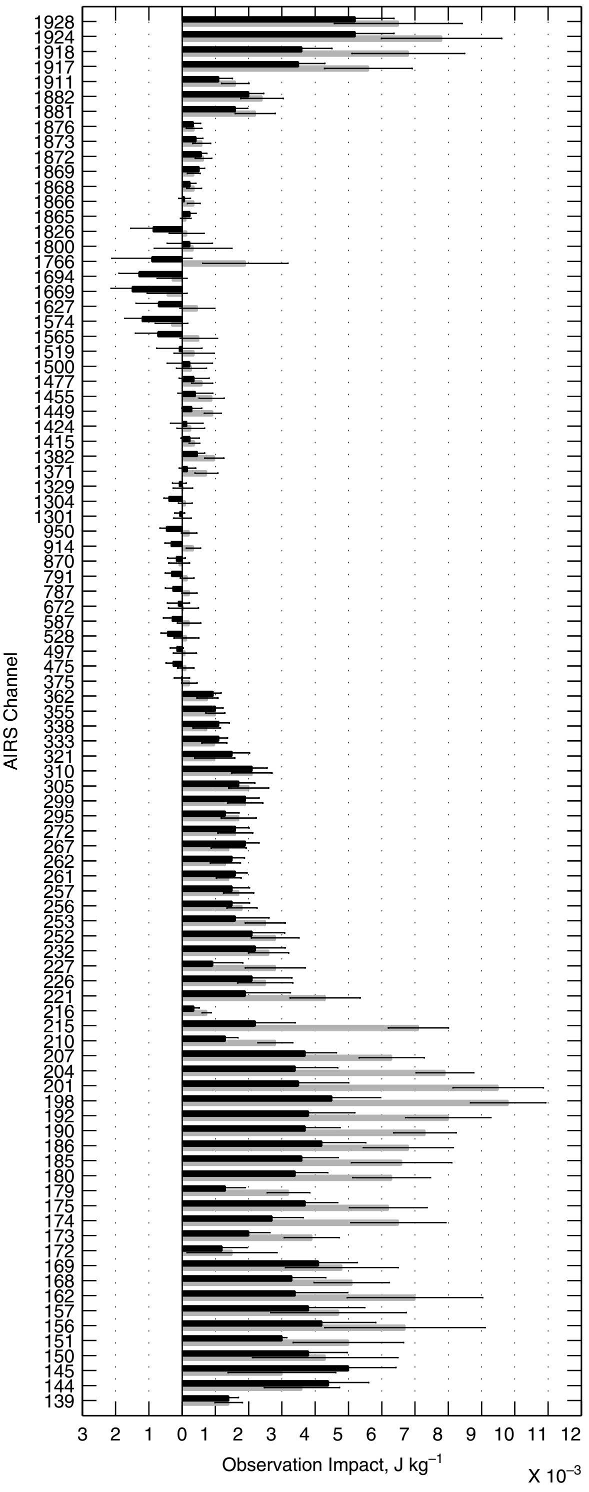

Figure 11. As Figure 10, but for tropospheric AIRS channels used in the GEOS-5 GSI.

The higher forecast skill in the OSSE might be expected to yield experimental results that underestimate the potential impact of a new observing system, as there is less available room for improvement in the forecasts. However, the difference in skill level of the five-day forecast is approximately 10% in terms of RMS error and 20% in terms of anomaly correlation error for real data; this discrepancy is not overly large and may still yield useful experimental results. The reduced observation impacts calculated with the adjoint are likely due to the improved forecast skill in the OSSE. As the relative ranking of observation impact between data types is generally preserved in the OSSE, it is anticipated that a new observing system could be evaluated qualitatively using the adjoint in the OSSE.

As demonstrated by the AMSU-A and AIRS results in section 3.2, matching observation impacts between the OSSE and the real world becomes more difficult when looking at individual satellite channels. It is also expected that achieving a good match will be more difficult when considering regional impacts or more difficult metrics such as precipitation. When evaluating a new observation type, this issue becomes even more challenging due to the lack of real observations for tuning the synthetic observation generator and observation-error characteristics. One way to attempt to mitigate this problem is to look at the performance of similar existing data types in comparison to real data to determine if the OSSE is capable of accurately representing the observation impact and behaviour.

Both the ability of the NR to accurately represent the behaviour of concern and the relative skill of the OSSE forecasts should be evaluated before performing experiments. New observing systems that observe at very high resolution, such as airborne Doppler radar, or phenomena that likewise require high model resolution to be accurately portrayed, such as tropical cyclone structure and intensity, are not suited to study with the present GMAO OSSE due to the low resolution of the present NR. Studies of features for which modelling skill is relatively low, such as precipitation or near-surface fields, should also be approached with caution.

The robustness of OSSE results can be estimated by performing multiple experiments that explore the response of the data assimilation system and forecasts to the new observations. By varying the observation-error magnitude and characteristics such as error correlation and by testing the new observations in multiple seasons and under different synoptic states, a better understanding of the impact of the new observational data can be gained. One significant advantage of the current generation of OSSEs is that the much longer NR in comparison to previous OSSEs (Becker *et al.*, 1996) allows for more thorough testing and exploration of new observations rather than case-studies. Although the behaviour of the GMAO OSSE is not a perfect analogue of the real world system, this OSSE has considerably more sophisticated representations of observations and observation error than previous OSSEs, and the GMAO OSSE has also been rigorously calibrated both to improve the performance and to understand the shortcomings of the OSSE system.

Although the NR and synthetic observation suite may lag behind operational modelling advances, an OSSE framework may be used for an extended period after development. For example, the OSSE framework described by Becker *et al.* (1996) was generated in the mid-1990s and has been used for experiments for more than a decade (e.g. Cardinali *et al.*, 1998; Stoffelen *et al.*, 2006; Masutani *et al.*, 2010). The methods and procedures for NR evaluation and generation of synthetic observations and their errors can be brought forward to shorten the development time for the future incarnations of the OSSE.

In addition to the traditional use of OSSEs for evaluating new observing systems, the OSSE can be used to explore the behaviour of data assimilation systems and numerical weather prediction models; both of these purposes are of interest for NASA GMAO. Updating the suite of synthetic observations in the GMAO OSSE to the 2011–2012 dataset is currently underway. Future investigations include quantifying the relative roles of model error and initial condition error in the evolution of forecast error and exploring the impact of observation error on analysis and forecast skill.

Acknowledgements

The ECMWF Nature Run was provided by Erik Andersson through arrangements made by Michiko Masutani. The authors wish to thank three anonymous reviewers for their comments that led to significant improvements in this manuscript. Support for this project was encouraged by Michele Rienecker and provided by grant numbers WBS 802678.02.17.01.06 and MAP/04-0000-0080 from the National Aeronautics and Space Administration and grant number ATM-0745906 from the National Science Foundation.

References

Becker BD, Roquet H, Stoffelen A. 1996. A simulated future atmospheric observation database including ATOVS, ASCAT, and DWL. *Bull. Am. Meteorol. Soc.* **77**: 2279–2294.

Buizza R. 1997. Potential forecast skill of ensemble prediction and spread and skill distributions of the ECMWF ensemble prediction system. *Mon. Weather Rev.* **125**: 99–119.

Cardinali C, Pailleux J, Thépaut J-N. 1998. 'Use of simulated Doppler wind lidar data in NWP: an impact study'. Groupe de Modelisation pour L'assimilation et le Prevision, Note de travail 6, Centre Nationale de la Recherche Scientifique, Météo-France: Toulouse.

Errico RM. 2000. Interpretations of the total energy and rotational energy norms applied to determination of singular vectors. *Tellus* **59A**: 273–276.

Errico RM, Yang R, Masutani M, Woollen JS. 2007. The estimation of analysis error characteristics using an observation systems simulation experiment. *Meteorol. Z.* **16**: 695–708.

Errico RM, Yang R, Privé N, Tai K-S, Todling R, Sienkiewicz M, Guo J. 2012. Validation of version one of the Observing System Simulation Experiments at the Global Modeling and Assimilation Office. *Q. J. R. Meteorol. Soc.* accepted

Gelaro R, Zhu Y. 2009. Examination of observation impacts derived from observing system experiments (OSEs) and adjoint models. *Tellus* **61A**: 179–193.

Han Y, van Delst P, Liu Q, Weng F, Yan B, Treadon R, Derber J. 2006. JCSDA Community Radiative Transfer Model (CRTM) – version 1. OAA Tech. Report 122.

Kleist DT, Parrish DF, Derber JC, Treadon R, Wu W-S, Lord S. 2009. Introduction of the GSI into the NCEP global data assimilation system. *Weather and Forecasting* **24**: 1691–1705.

Mann HB, Whitney DR. 1947. On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Statist.* **18**: 50–60.

Masutani M, Woollen JS, Lord SJ, Kleepsies TJ, Emmitt GD, Sun H, Wood SA, Greco S, Terry J, Treadon RE, Campana KA. 2006. 'Observing System Simulation Experiments and NCEP'. Office Note 451. National Centers for Environmental Prediction: Camp Springs, Maryland.

Masutani M, Andersson E, Terry J, Reale O, Jusem JC, Riishøjgaard LP, Schlatter T, Stoffelen A, Woollen JS, Lord S, Toth Z, Song Y, Kleist D, Xie Y, Privé N, Liu E, Sun H, Emmit D, Greco S, Wood SA, Marseille G-J, Errico RM, Yang R, McConaughy G, Devenji D, Weygandt S, Tompkins A, Jung T, Anantharaj V, Hill C, Fitzpatrick P, Weng F, Zhu T, Boukabara S. 2007. 'Progress in Joint OSSEs'. In *Preprint volume for 18th Conference on Numerical Weather Prediction*, 25–29 June 2007 Park City, UT. Amer. Meteorol. Soc: Boston.

Masutani M, Woollen JS, Lord SJ, Emmitt GD, Kleespies TJ, Wood SA, Greco S, Sun H, Terry J, Kapoor V, Treadon RE, Campana KA. 2010. Observing system simulation experiments at the National Centers for Environmental Prediction. *J. Geophys. Res.* **115**: D07101, DOI: 10.1029/2009JD012528

McCarty W, Errico RM, Gelaro R. 2012. Cloud coverage in the Joint OSSE Nature Run. *Mon. Weather Rev.* **140**: 1863–1871.

Reale O, Terry J, Masutani M, Andersson E, Riishøjgaard LP, Jusem JC. 2007. Preliminary evaluation of the European Centre for Medium-range Weather Forecasts (ECMWF) nature run over the tropical Atlantic and African monsoon region. *Geophys. Res. Lett.* **34**: L22810, DOI: 10.1029/2007GL031640

Rienecker MM, Suarez MJ, Todling R, Bacmeister J, Takacs L, Liu H-C, Gu W, Sienkiewicz M, Koster RD, Gelaro R, Stajner I, Nielsen JE. 2008. 'The GEOS-5 data assimilation system – documentation of versions 5.0.1, 5.1.0 and 5.2.0'. Technical Report Series on Global Modeling and Data Assimilation No. 27. NASA, Goddard Space Flight Center: Greenbelt, Maryland.

Rosenblueth A, Wiener N. 1945. The role of models in science. *Phil. Sci.* **12**: 316–321.

Stoffelen A, Marseille GJ, Bouttier F, Vasiljevic D, de Haan S, Cardinali C. 2006. ADM–Aeolus Doppler wind lidar Observing System Simulation Experiment. *Q. J. R. Meteorol. Soc.* **132**: 1927–1947.

Talagrand O. 1981. A study of the dynamics of four-dimensional data assimilation. *Tellus* **33**: 43–60.

Zapotocny TH, Jung JA, Le Marshall JF, Treadon RE. 2008. A two-season impact study of four satellite data types and rawinsonde data in the NCEP global data assimilation system. *Weather and Forecasting* **23**: 80–100.

Zhu Y, Gelaro R. 2008. Observation sensitivity calculations using the adjoint of the Gridpoint Statistical Interpolation (GSI) analysis system. *Mon. Weather Rev.* **136**: 335–351.